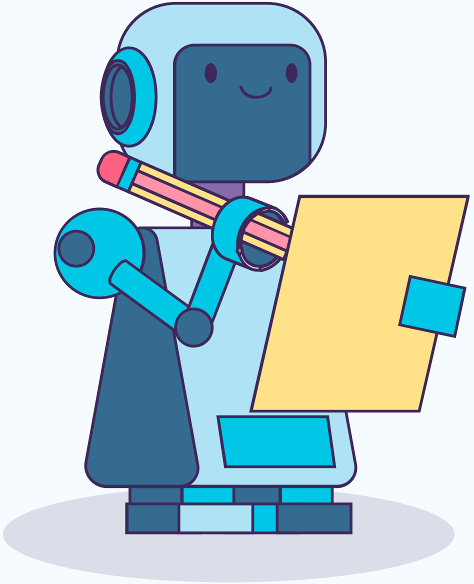
Hadoop WordCount Comparison

Kelompok 6



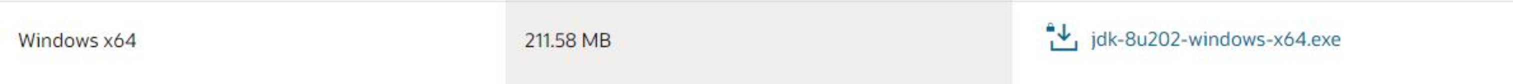
Anggota

Muhammad Farrel Mirawan - 2106731554
Akmal Rabbani - 2106731610
Arka Brian Dewara - 2106731421
Nevanda Fairuz Pahlevi - 2106731541

01

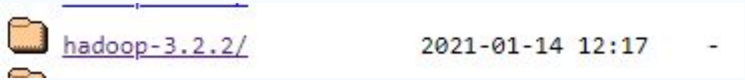
Hadoop

Download Hadoop Prerequisite



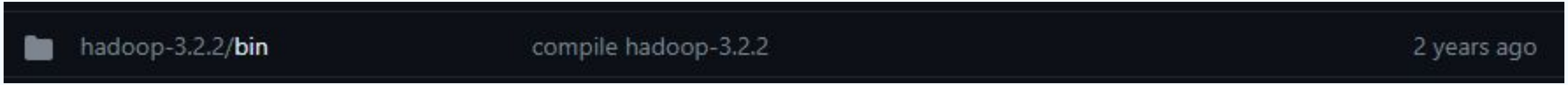
Windows x64	211.58 MB	jdk-8u202-windows-x64.exe

JAVA 8 : https://www.oracle.com/id/java/technologies/javase/javase8-archive-downloads.html



```
hadoop-3.2.2/            2021-01-14 12:17    -
```

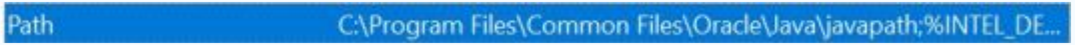
Hadoop : https://archive.apache.org/dist/hadoop/common/ kemudian extract



hadoop-3.2.2/**bin**	compile hadoop-3.2.2	2 years ago

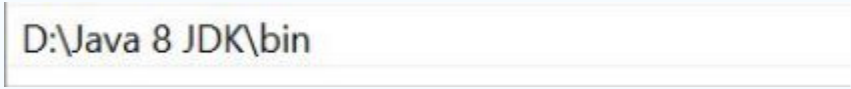
Additional Files : https://github.com/cdarlint/winutils

Konfigurasi Path Variable Java



Path C:\Program Files\Common Files\Oracle\Java\javapath;%INTEL_DE...

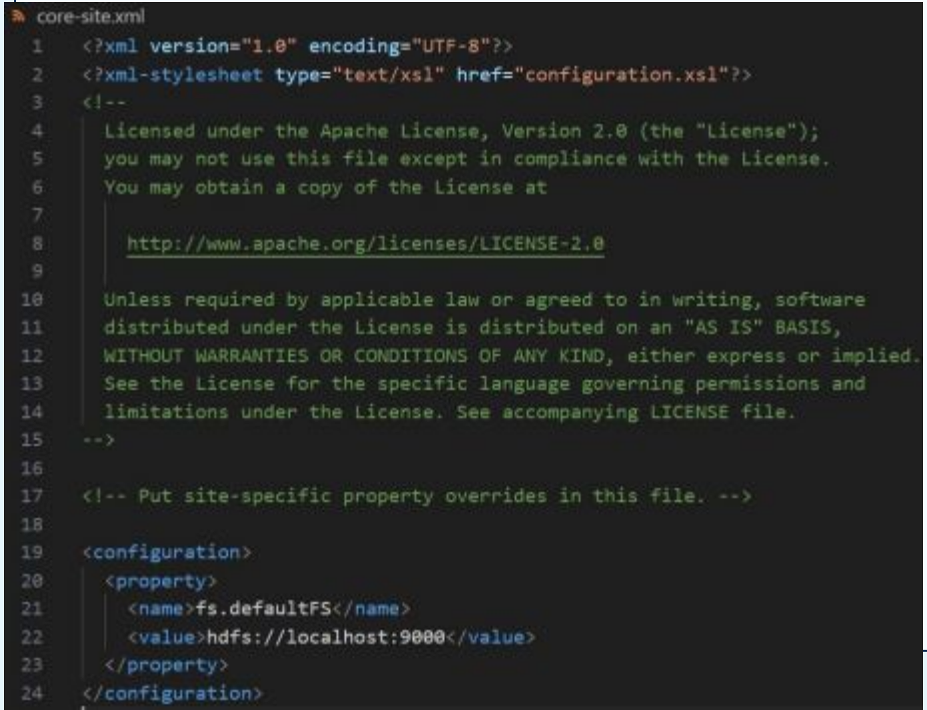
Klik edit kemudian tambahkan path seperti dibawah



D:\Java 8 JDK\bin

Konfigurasi Hadoop

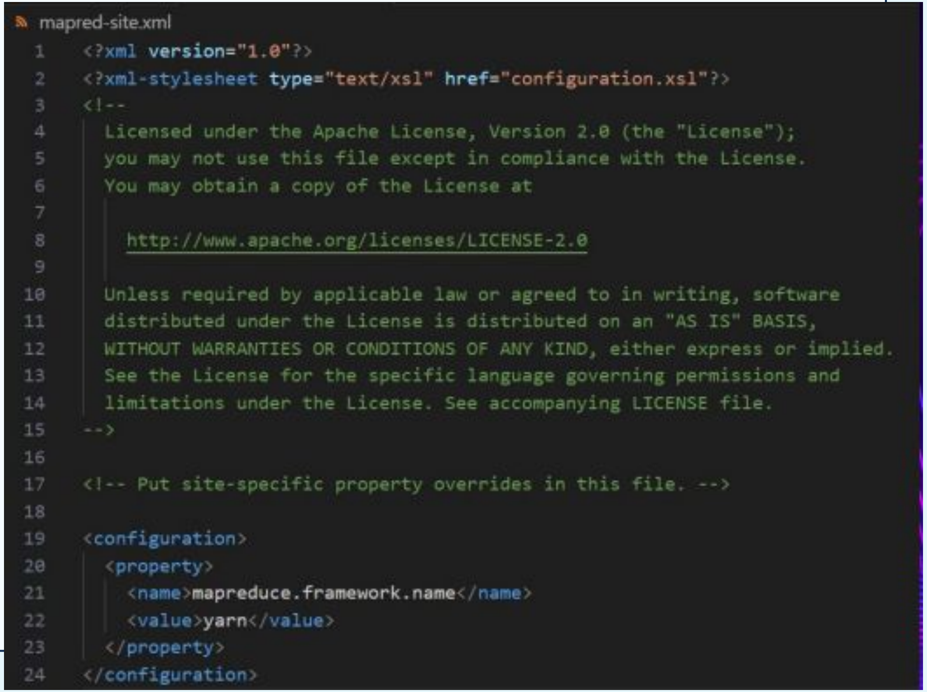
Core-Site.xml



core-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

Mapred-Site.xml

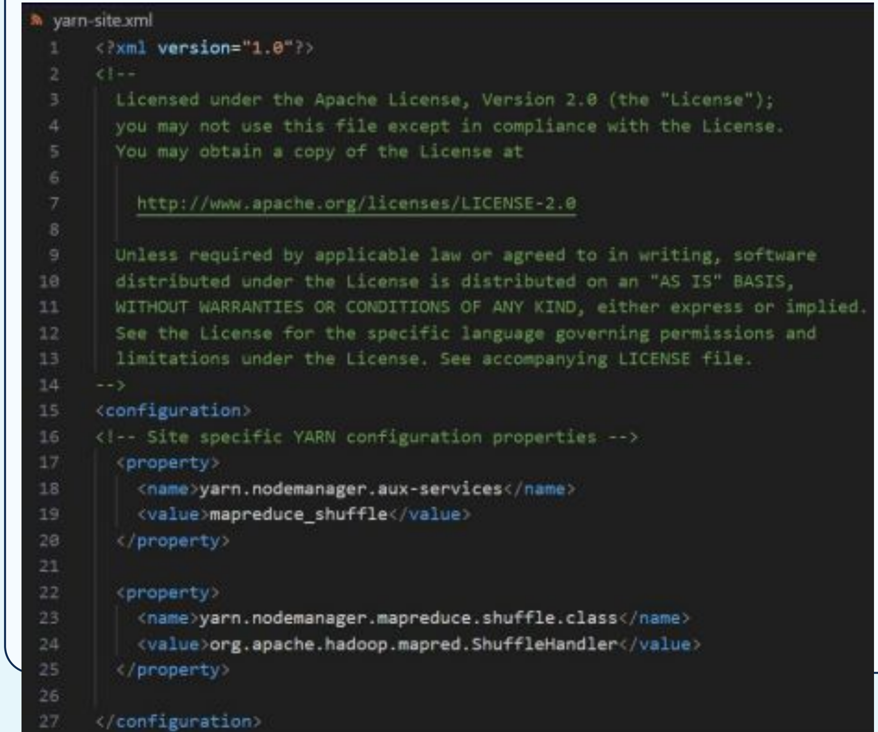


mapred-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

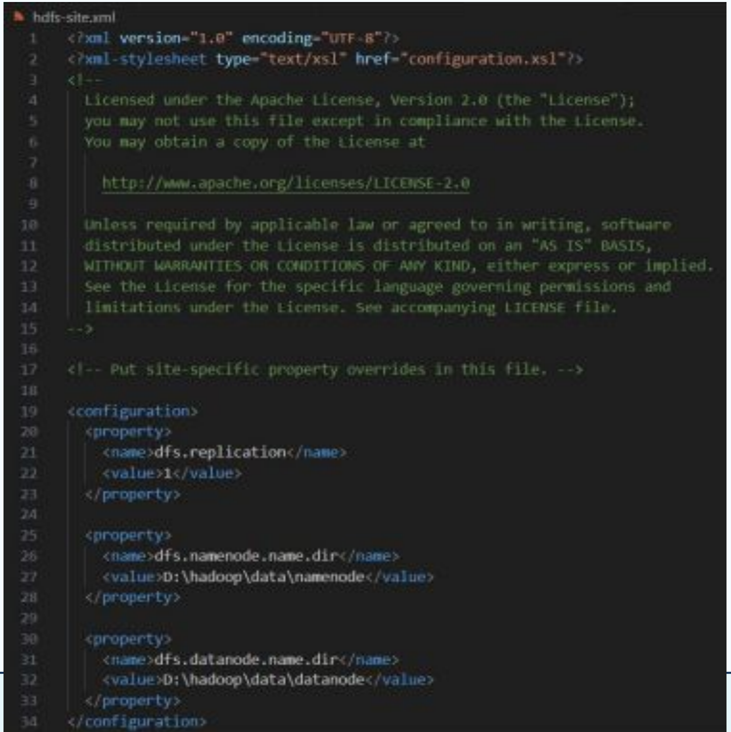
Konfigurasi Hadoop

Yarn-Site.xml



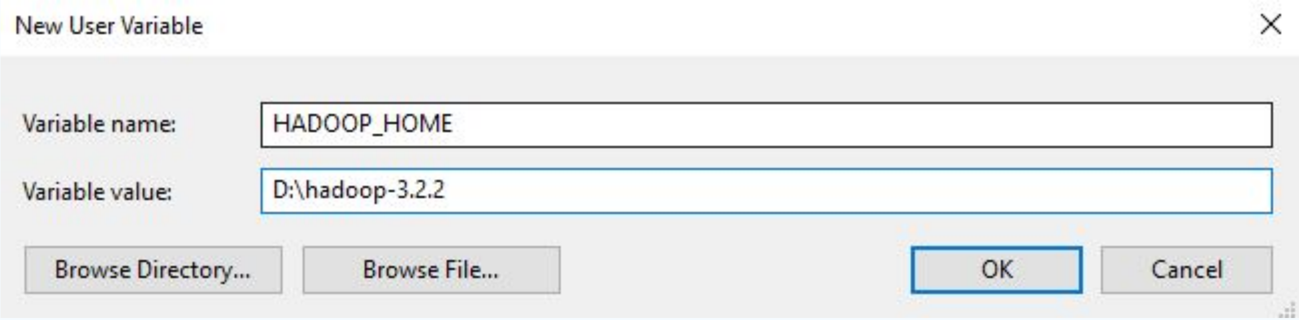
```xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
<!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>

  <property>
    <name>yarn.nodemanager.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>

</configuration>
```

hdfs-Site.xml



```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>

  <property>
    <name>dfs.namenode.name.dir</name>
    <value>D:\hadoop\data\namenode</value>
  </property>

  <property>
    <name>dfs.datanode.name.dir</name>
    <value>D:\hadoop\data\datanode</value>
  </property>
</configuration>
```

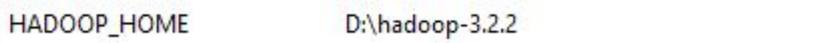
Membuat variable hadoop



New User Variable

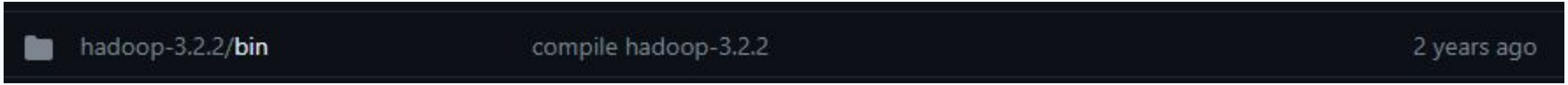
Variable name: HADOOP_HOME

Variable value: D:\hadoop-3.2.2



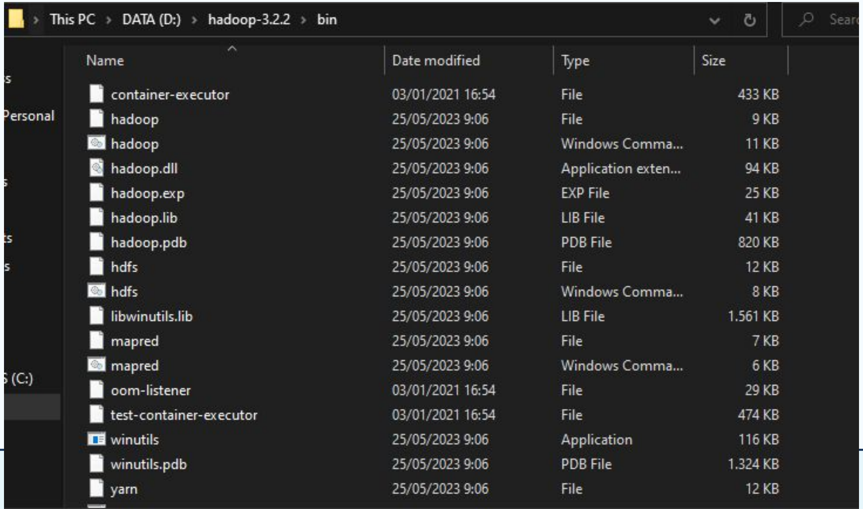
HADOOP_HOME	D:\hadoop-3.2.2

Download additional files dan masukkan ke /bin



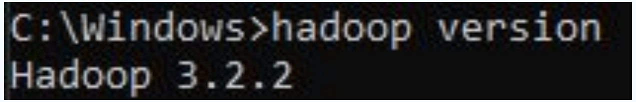
hadoop-3.2.2/bin compile hadoop-3.2.2 2 years ago

Additional Files : https://github.com/cdarlint/winutils

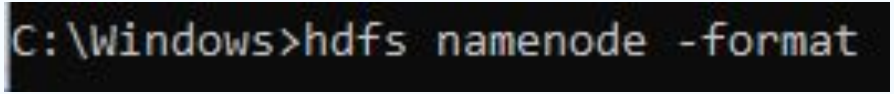


Name	Date modified	Type	Size
container-executor	03/01/2021 16:54	File	433 KB
hadoop	25/05/2023 9:06	File	9 KB
hadoop	25/05/2023 9:06	Windows Comma...	11 KB
hadoop.dll	25/05/2023 9:06	Application exten...	94 KB
hadoop.exp	25/05/2023 9:06	EXP File	25 KB
hadoop.lib	25/05/2023 9:06	LIB File	41 KB
hadoop.pdb	25/05/2023 9:06	PDB File	820 KB
hdfs	25/05/2023 9:06	File	12 KB
hdfs	25/05/2023 9:06	Windows Comma...	8 KB
libwinutils.lib	25/05/2023 9:06	LIB File	1.561 KB
mapred	25/05/2023 9:06	File	7 KB
mapred	25/05/2023 9:06	Windows Comma...	6 KB
oom-listener	03/01/2021 16:54	File	29 KB
test-container-executor	03/01/2021 16:54	File	474 KB
winutils	25/05/2023 9:06	Application	116 KB
winutils.pdb	25/05/2023 9:06	PDB File	1.324 KB
yarn	25/05/2023 9:06	File	12 KB

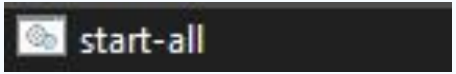
Verifikasi Installasi Hadoop



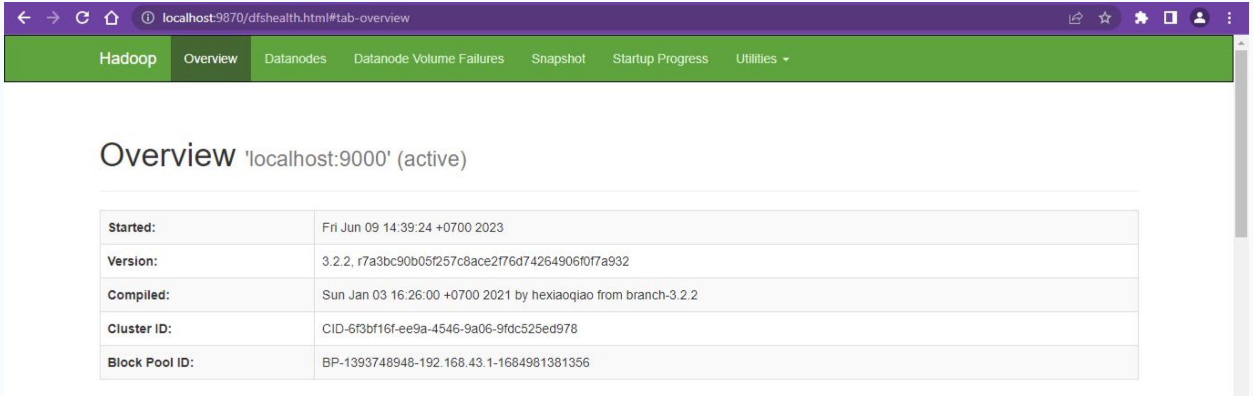
```
C:\Windows>hadoop version
Hadoop 3.2.2
```



```
C:\Windows>hdfs namenode -format
```

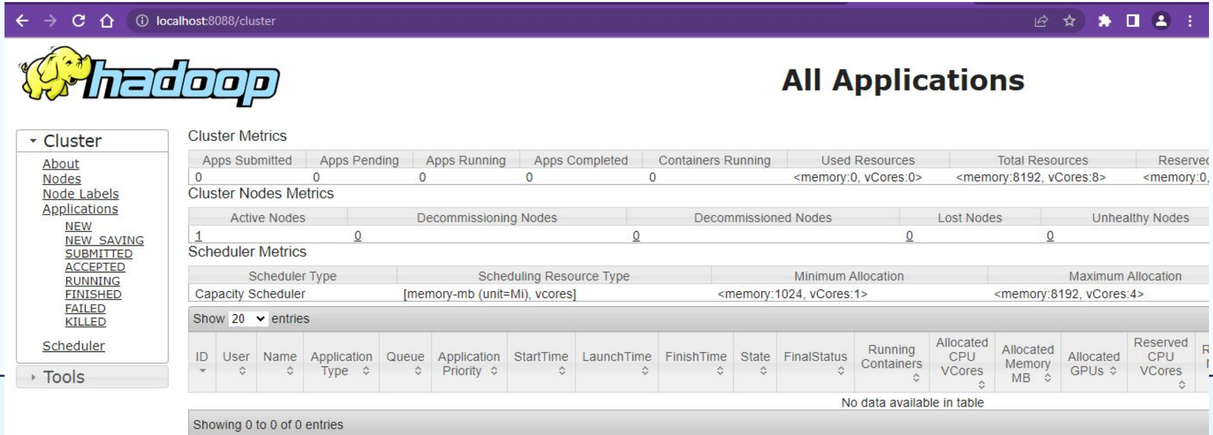


start-all



Overview 'localhost:9000' (active)

Started:	Fri Jun 09 14:39:24 +0700 2023
Version:	3.2.2, r7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled:	Sun Jan 03 16:26:00 +0700 2021 by hexiaoqiao from branch-3.2.2
Cluster ID:	CID-6f3bf16f-ee9a-4546-9a06-9fdc525ed978
Block Pool ID:	BP-1393748948-192.168.43.1-1684981381356



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reservec
0	0	0	0	0	<memory:0, vCores:0>	<memory:8192, vCores:8>	<memory:0,

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes
1	0	0	0	0

Scheduler Metrics

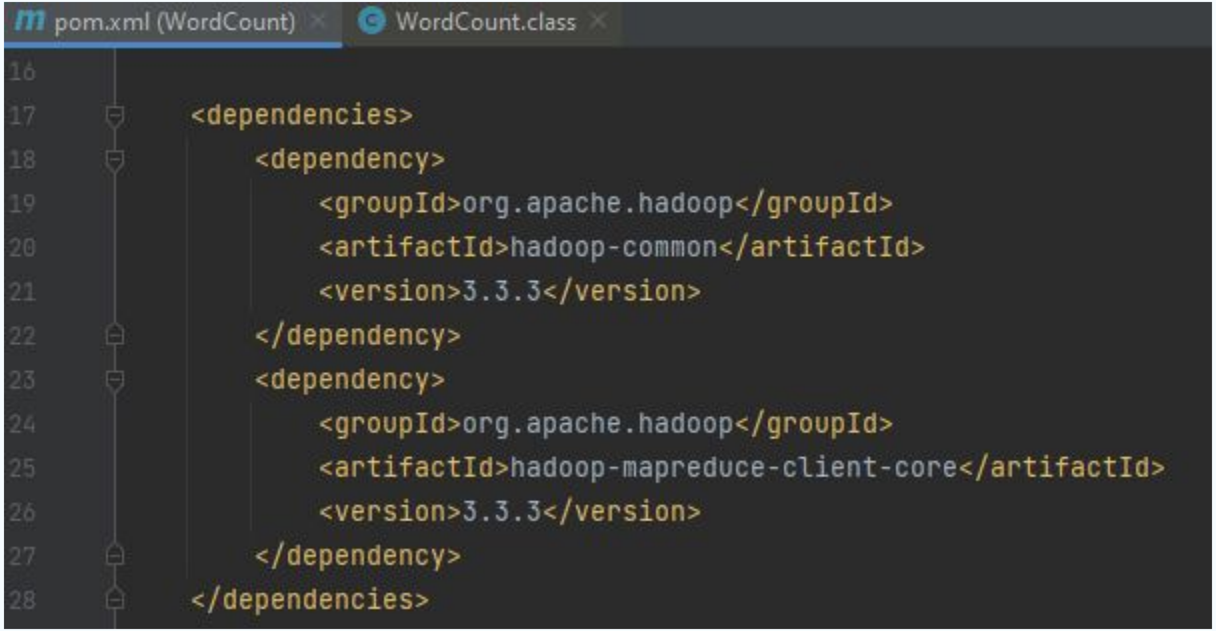
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	R
No data available in table																

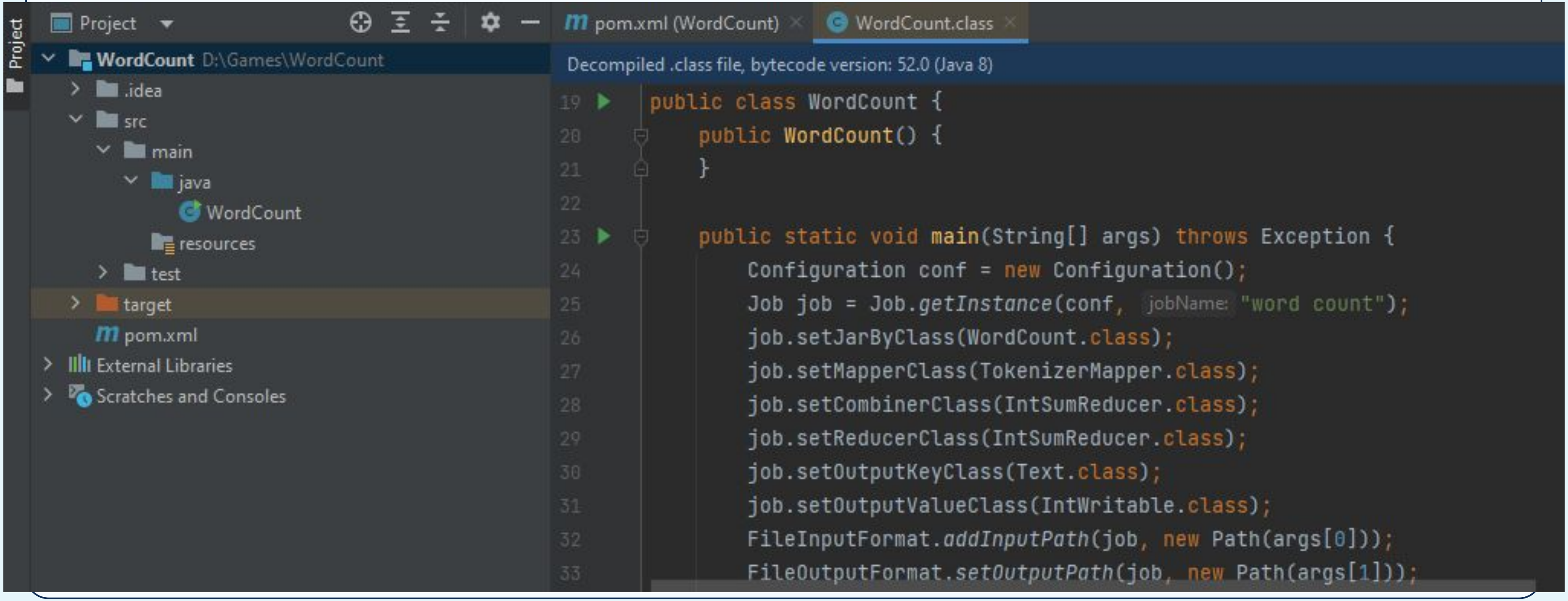
Showing 0 to 0 of 0 entries

WordCount Setup in Hadoop



```xml

    <dependencies>
        <dependency>
            <groupId>org.apache.hadoop</groupId>
            <artifactId>hadoop-common</artifactId>
            <version>3.3.3</version>
        </dependency>
        <dependency>
            <groupId>org.apache.hadoop</groupId>
            <artifactId>hadoop-mapreduce-client-core</artifactId>
            <version>3.3.3</version>
        </dependency>
    </dependencies>
```

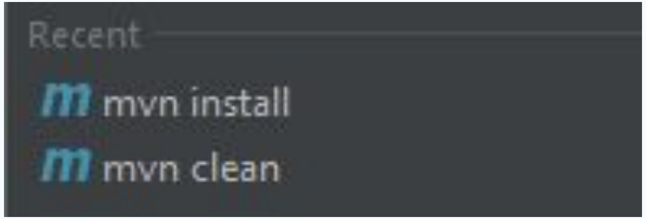
WordCount Setup in Hadoop



Decompiled .class file, bytecode version: 52.0 (Java 8)

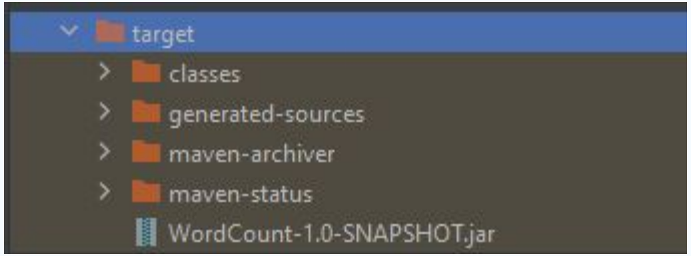
```java
public class WordCount {
    public WordCount() {
    }

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, jobName: "word count");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(TokenizerMapper.class);
        job.setCombinerClass(IntSumReducer.class);
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
```

WordCount Setup in Hadoop



```
mvn install
mvn clean
```

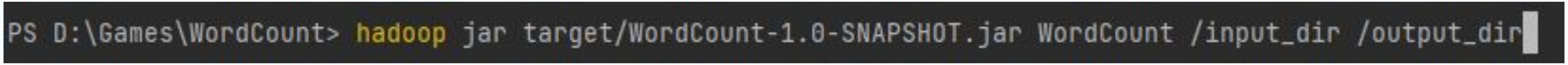
Create Jar using maven



```
target
  classes
  generated-sources
  maven-archiver
  maven-status
  WordCount-1.0-SNAPSHOT.jar
```

hadoop fs -mkdir /input_dir	Membuat folder input pada HDFS
hadoop fs -put /path/to/file.txt /input_dir	Meletakkan file text ke folder input HDFS

Menjalankan jar

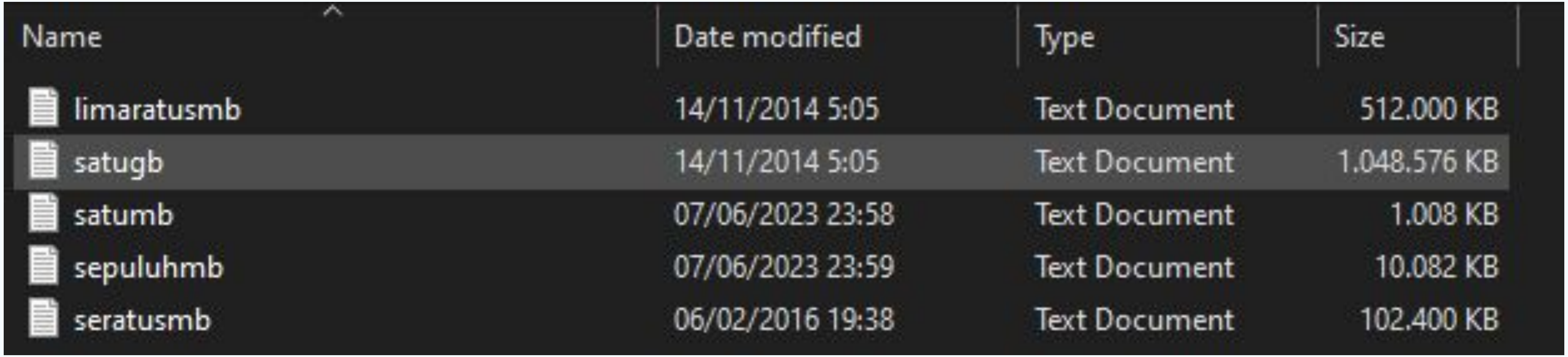


```
PS D:\Games\WordCount> hadoop jar target/WordCount-1.0-SNAPSHOT.jar WordCount /input_dir /output_dir
```

02

Hadoop Vs Java

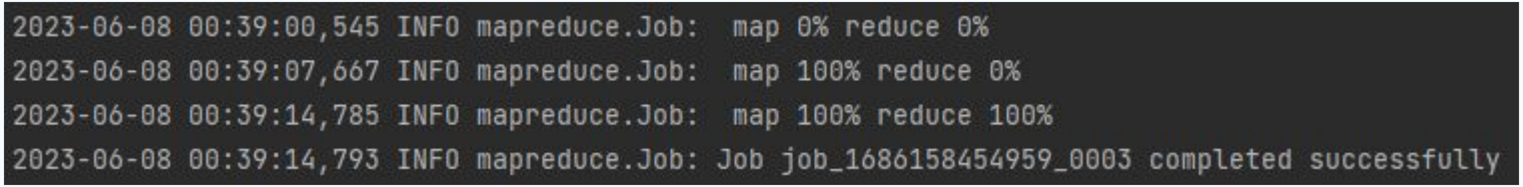
File yang akan diuji



Name	Date modified	Type	Size
limaratusmb	14/11/2014 5:05	Text Document	512.000 KB
satugb	14/11/2014 5:05	Text Document	1.048.576 KB
satumb	07/06/2023 23:58	Text Document	1.008 KB
sepuluhmb	07/06/2023 23:59	Text Document	10.082 KB
seratusmb	06/02/2016 19:38	Text Document	102.400 KB

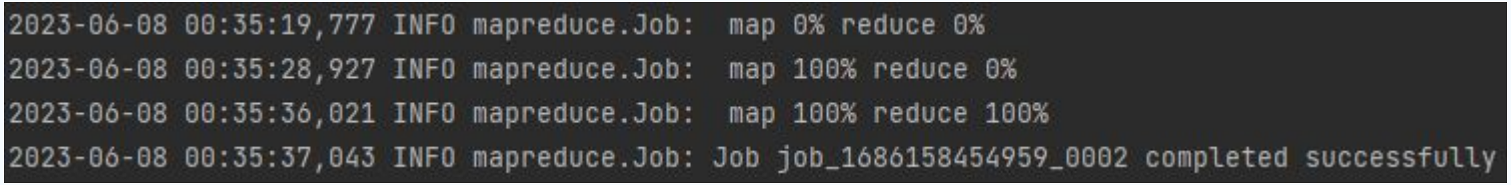
Hadoop

1 Mb = 14.248 s



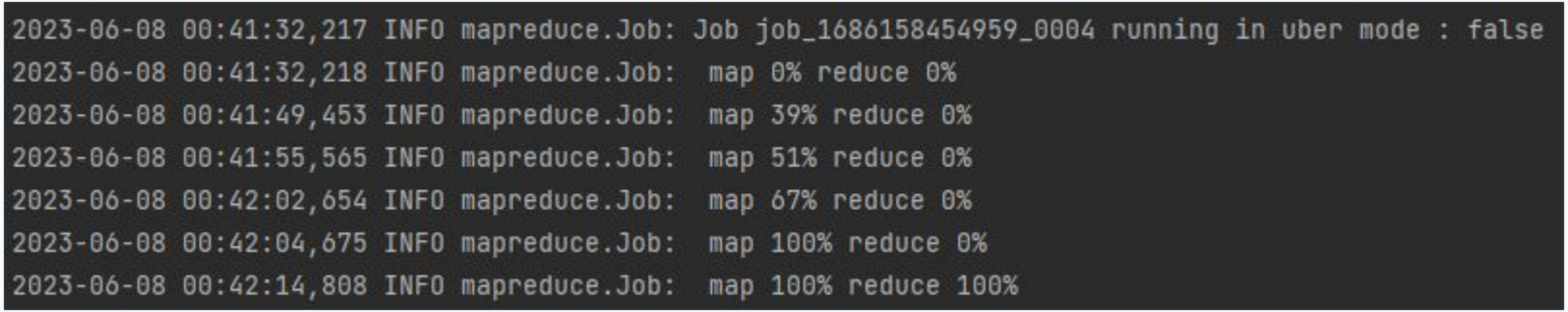
```
2023-06-08 00:39:00,545 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 00:39:07,667 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 00:39:14,785 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 00:39:14,793 INFO mapreduce.Job: Job job_1686158454959_0003 completed successfully
```

10 Mb = 17.266 s



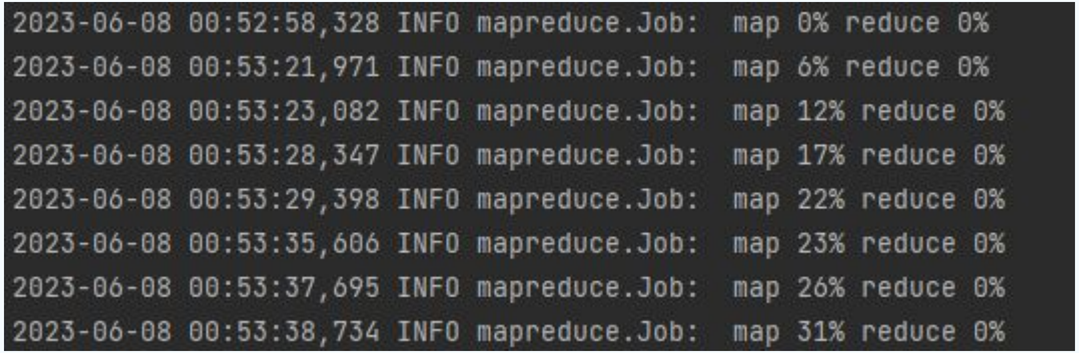
```
2023-06-08 00:35:19,777 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 00:35:28,927 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 00:35:36,021 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 00:35:37,043 INFO mapreduce.Job: Job job_1686158454959_0002 completed successfully
```

100 Mb = 42.591 s

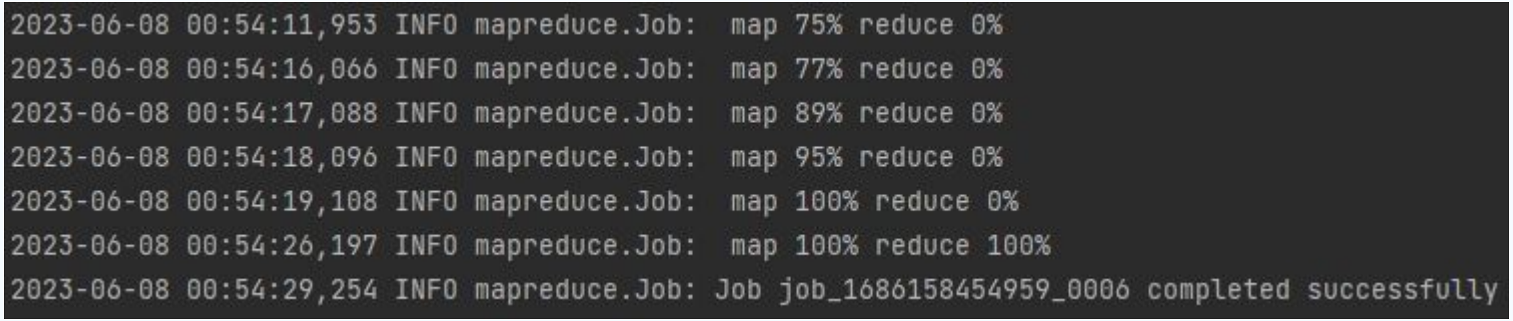


```
2023-06-08 00:41:32,217 INFO mapreduce.Job: Job job_1686158454959_0004 running in uber mode : false
2023-06-08 00:41:32,218 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 00:41:49,453 INFO mapreduce.Job:  map 39% reduce 0%
2023-06-08 00:41:55,565 INFO mapreduce.Job:  map 51% reduce 0%
2023-06-08 00:42:02,654 INFO mapreduce.Job:  map 67% reduce 0%
2023-06-08 00:42:04,675 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 00:42:14,808 INFO mapreduce.Job:  map 100% reduce 100%
```

500 Mb = 91.074 s

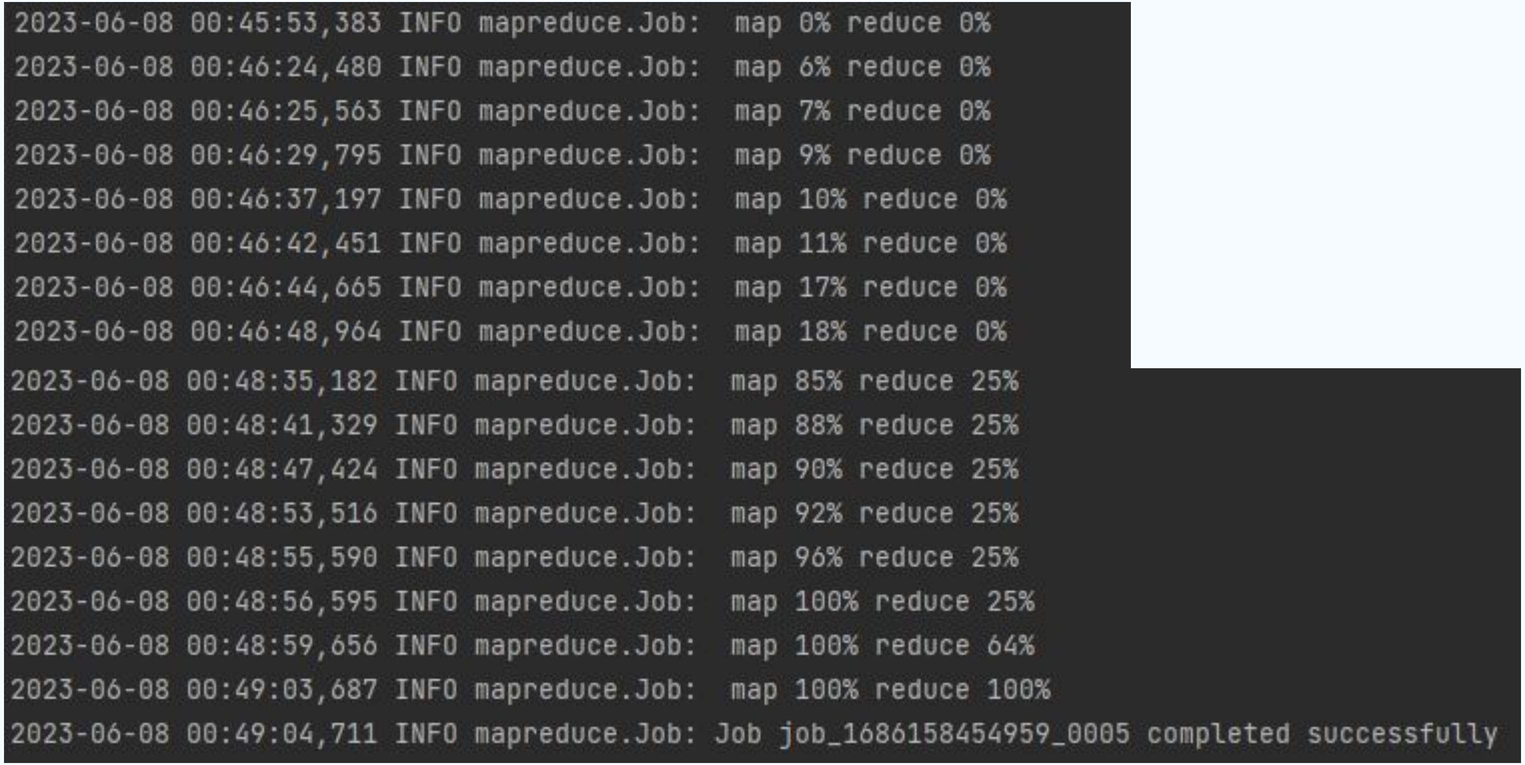


```
2023-06-08 00:52:58,328 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 00:53:21,971 INFO mapreduce.Job:  map 6% reduce 0%
2023-06-08 00:53:23,082 INFO mapreduce.Job:  map 12% reduce 0%
2023-06-08 00:53:28,347 INFO mapreduce.Job:  map 17% reduce 0%
2023-06-08 00:53:29,398 INFO mapreduce.Job:  map 22% reduce 0%
2023-06-08 00:53:35,606 INFO mapreduce.Job:  map 23% reduce 0%
2023-06-08 00:53:37,695 INFO mapreduce.Job:  map 26% reduce 0%
2023-06-08 00:53:38,734 INFO mapreduce.Job:  map 31% reduce 0%
```



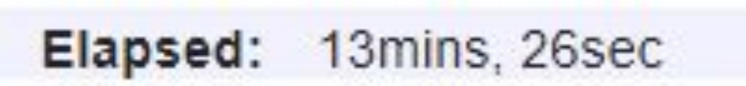
```
2023-06-08 00:54:11,953 INFO mapreduce.Job:  map 75% reduce 0%
2023-06-08 00:54:16,066 INFO mapreduce.Job:  map 77% reduce 0%
2023-06-08 00:54:17,088 INFO mapreduce.Job:  map 89% reduce 0%
2023-06-08 00:54:18,096 INFO mapreduce.Job:  map 95% reduce 0%
2023-06-08 00:54:19,108 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 00:54:26,197 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 00:54:29,254 INFO mapreduce.Job: Job job_1686158454959_0006 completed successfully
```

1 Gb = 191.408 s



```
2023-06-08 00:45:53,383 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 00:46:24,480 INFO mapreduce.Job:  map 6% reduce 0%
2023-06-08 00:46:25,563 INFO mapreduce.Job:  map 7% reduce 0%
2023-06-08 00:46:29,795 INFO mapreduce.Job:  map 9% reduce 0%
2023-06-08 00:46:37,197 INFO mapreduce.Job:  map 10% reduce 0%
2023-06-08 00:46:42,451 INFO mapreduce.Job:  map 11% reduce 0%
2023-06-08 00:46:44,665 INFO mapreduce.Job:  map 17% reduce 0%
2023-06-08 00:46:48,964 INFO mapreduce.Job:  map 18% reduce 0%
2023-06-08 00:48:35,182 INFO mapreduce.Job:  map 85% reduce 25%
2023-06-08 00:48:41,329 INFO mapreduce.Job:  map 88% reduce 25%
2023-06-08 00:48:47,424 INFO mapreduce.Job:  map 90% reduce 25%
2023-06-08 00:48:53,516 INFO mapreduce.Job:  map 92% reduce 25%
2023-06-08 00:48:55,590 INFO mapreduce.Job:  map 96% reduce 25%
2023-06-08 00:48:56,595 INFO mapreduce.Job:  map 100% reduce 25%
2023-06-08 00:48:59,656 INFO mapreduce.Job:  map 100% reduce 64%
2023-06-08 00:49:03,687 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 00:49:04,711 INFO mapreduce.Job: Job job_1686158454959_0005 completed successfully
```

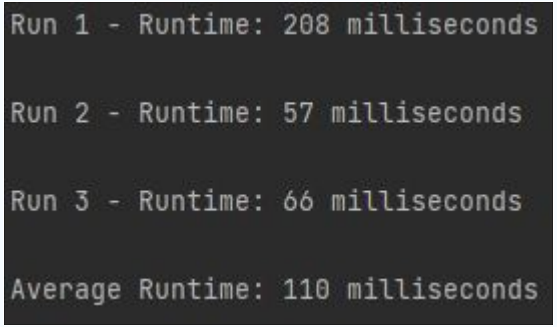
10 Gb = 806 s



Elapsed: 13mins, 26sec

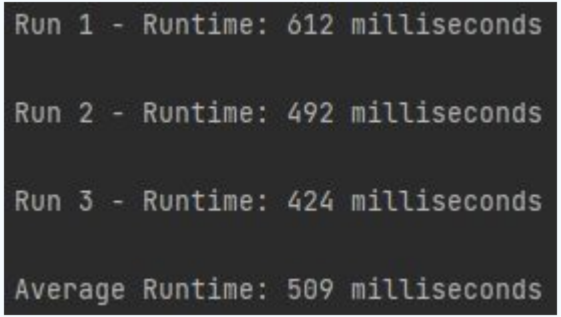
Java

1 Mb = 110 ms



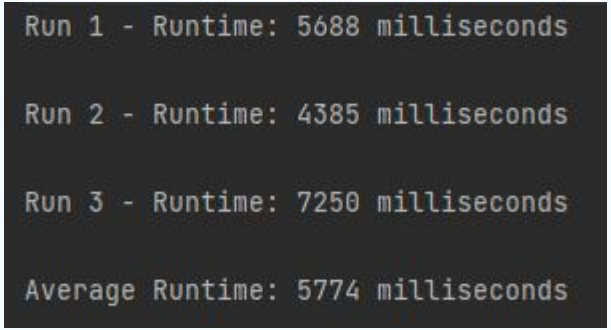
```
Run 1 - Runtime: 208 milliseconds

Run 2 - Runtime: 57 milliseconds

Run 3 - Runtime: 66 milliseconds

Average Runtime: 110 milliseconds
```

10 Mb = 509 ms



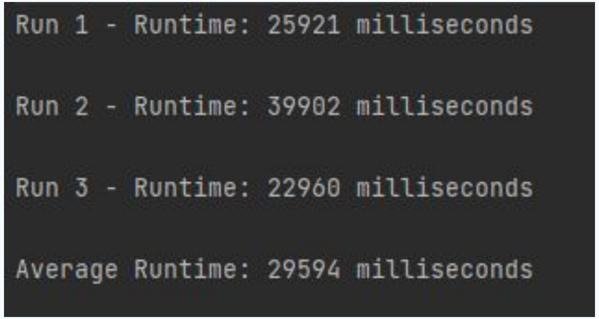
```
Run 1 - Runtime: 612 milliseconds

Run 2 - Runtime: 492 milliseconds

Run 3 - Runtime: 424 milliseconds

Average Runtime: 509 milliseconds
```

100 Mb = 5774 ms



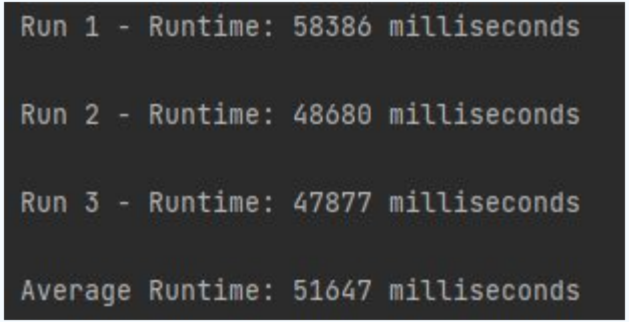
```
Run 1 - Runtime: 5688 milliseconds

Run 2 - Runtime: 4385 milliseconds

Run 3 - Runtime: 7250 milliseconds

Average Runtime: 5774 milliseconds
```

500 Mb = 29594 ms



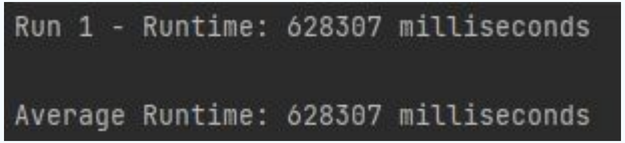
```
Run 1 - Runtime: 25921 milliseconds

Run 2 - Runtime: 39902 milliseconds

Run 3 - Runtime: 22960 milliseconds

Average Runtime: 29594 milliseconds
```

1 Gb = 51647 ms

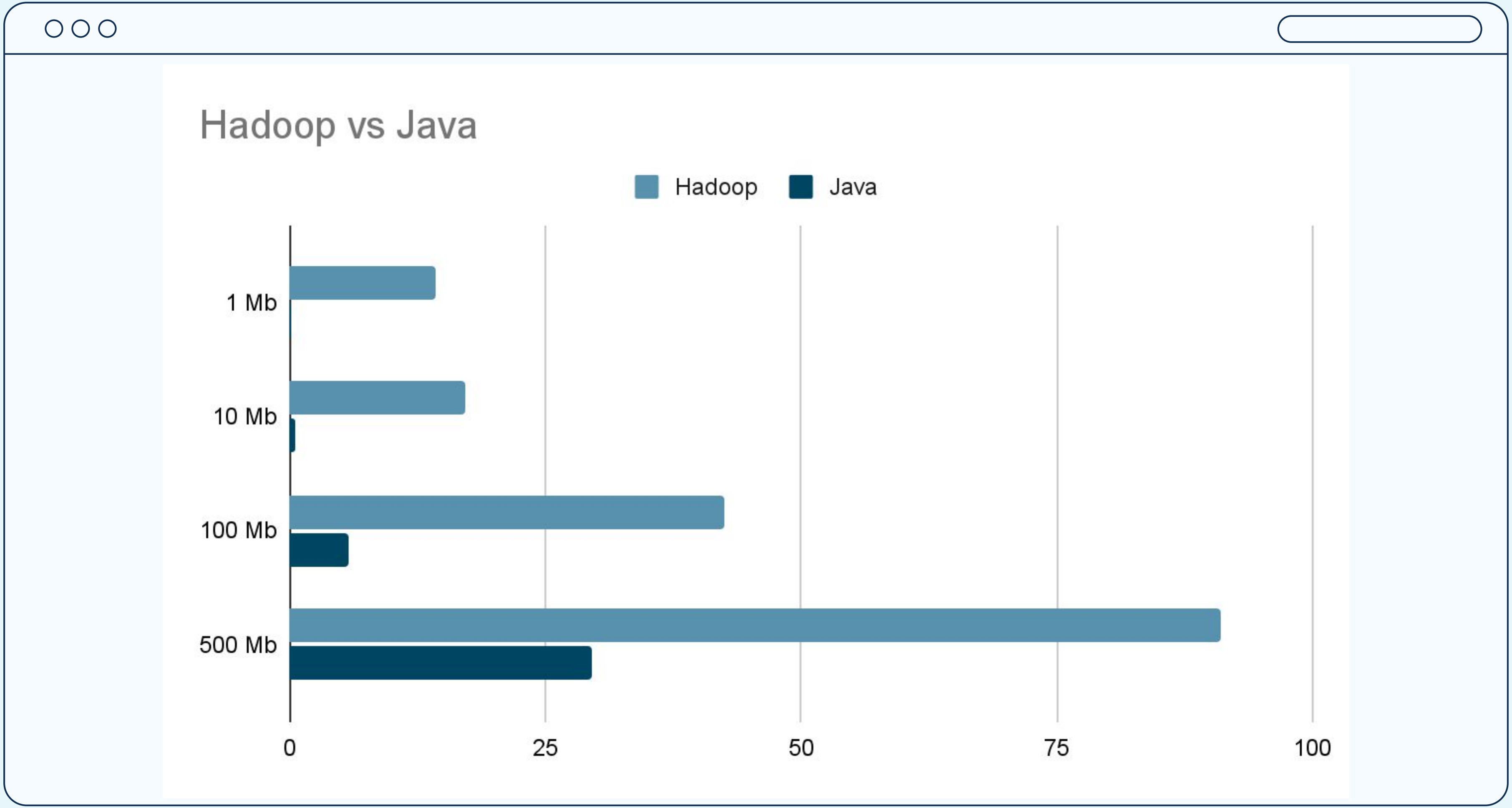


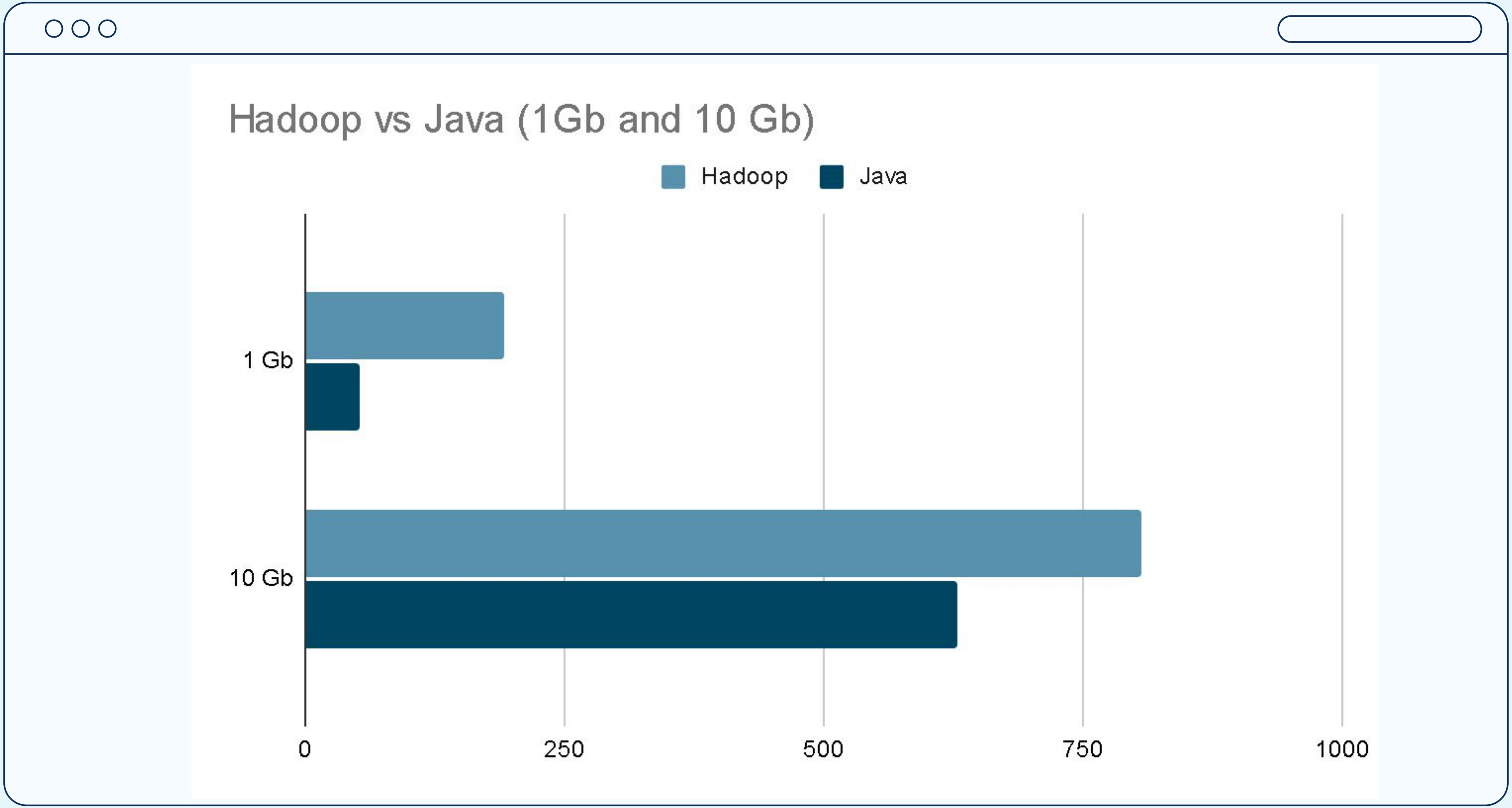
```
Run 1 - Runtime: 58386 milliseconds

Run 2 - Runtime: 48680 milliseconds

Run 3 - Runtime: 47877 milliseconds

Average Runtime: 51647 milliseconds
```

10 Gb = 628307 ms



```
Run 1 - Runtime: 628307 milliseconds

Average Runtime: 628307 milliseconds
```


Hadoop vs Java
Hadoop
Java
1 Mb
10 Mb
100 Mb
500 Mb
0
25
50
75
100


Hadoop vs Java (1Gb and 10 Gb)
Hadoop
Java
1 Gb
10 Gb
0
250
500
750
1000

Analisis

Percobaan perbandingan antara kecepatan program worcount dengan menggunakan hadoop mapreduce dan java sudah dilakukan. Yang dimana Java lebih cepat daripada Hadoop dalam kebanyakan kasus, kecuali untuk file 10GB di mana perbedaan kecepatan tidak signifikan. Faktor nya adalah sebagai berikut :

- Overhead Hadoop Hadoop memiliki beberapa overhead tambahan yang harus diatasi dalam lingkungan yang didistribusikan.
- Latensi jaringan Dalam Hadoop, tugas-tugas pemetaan (mapping) dan pengurutan (reducing) dijalankan pada beberapa mesin dalam sebuah cluster.
- Didesain untuk big data Hadoop dirancang untuk menangani pemrosesan data yang sangat besar. Dalam kasus-kasus di mana ukuran file yang diuji cukup besar (diatas 10GB), Hadoop dapat menghasilkan hasil runtime yang lebih cepat dibanding dengan Java biasa..

Kesimpulan

Dari percobaan yang telah dilakukan didapat kesimpulan bahwa Hadoop MapReduce lebih lambat daripada program Java biasa dalam pengolahan data berukuran kecil (dibawah 10GB) karena overhead Hadoop dan latensi jaringan. Namun, saat ukuran file menjadi sangat besar (diatas 10GB), Hadoop dapat memberikan keuntungan dalam hal skalabilitas dan memanfaatkan arsitektur distribusi untuk memproses data dengan efisien sehingga akan lebih cepat dibanding Java pada data yang sangat besar.

THANK YOU